DOCUMENT RESUME

ED 340 736 TM 017 793

AUTHOR Connell, Peggy H.
TITLE An Analysis of Aptitude-Achievement Discrepancy Formulas in Learning Disability Assessment.
PUB DATE Nov 91
NOTE 25p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (20th, Lexington, KY, November 12-15, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Analysis of Variance; Aptitude Tests; Comparative Testing; *Educational Diagnosis; Elementary Education; *Elementary School Students; Estimation (Mathematics); *Intelligence Tests; *Learning Disabilities; Models; Regression (Statistics); Scores; *Scoring Formulas
IDENTIFIERS *Discrepancy Analysis; Kaufman Test of Educational Achievement; Wechsler Intelligence Scale for Children (Revised)

ABSTRACT

Results of applying different aptitude-achievement discrepancy models for test scores were compared for 150 second, 150 third, 150 fourth, and 130 fifth graders in several school systems in the Southeast who had been referred for psychoeducational testing for learning disabilities to determine whether different groups of students were identified as having a severe discrepancy according to the formula applied. Subjects included 200, 200, and 180 students of low, average, and high aptitude, respectively. Instruments used were the Wechsler Intelligence Scale for Children-Revised (WISC-R) and the Kaufman Test of Educational Achievement (K-TEA). The three models were: (1) simple difference; (2) estimated true difference score; and (3) frequency of regression prediction. A factorial two between- and one within-subjects with repeated measure analysis of variance design was used. The independent variables grade level and aptitude level crossed with the independent variable, discrepancy model. Each set of scores (WISC-R and K-TEA) was tested by each model, and discrepancies were computed. The lack of agreement between the regression model and others suggests that the model chosen could have a severe impact on the identification process. The simple difference and estimated true difference models could be used interchangeably. Implications for identification of aptitude-achievement discrepancies are discussed. Three data tables and a 40-item list of references are included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED340736

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[✓] This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Peggy H. Connell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Analysis of Aptitude-Achievement Discrepancy Formulas in Learning Disability Assessment

Peggy H. Connell

Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, Kentucky, November 13-15, 1991.

TM017793

BEST COPY AVAILABLE

Introduction

Since the inception of the definition of learning disability by the federal government in 1968, controversy has constantly plagued this area of special education. The debate has focused on the practical problems of definition and operational criteria for identification practices (Algozzine & Ysseldyke, 1983-84). The controversy has heightened, though, since the passage of Public Law 94-142, the Education for All Handicapped Children Act of 1975, which included the learning disabled among those considered handicapped. After the passage of this act the number of students identified as learning disabled soared. Clarizio and Phillips (1986b) stated that learning disabilities have now become the prevalent handicap of school-age children, and Clarizio and Phillips (1986a) have estimated that almost one million students are identified as learning disabled. Numerous researchers (e.g., Frankenberger & Harper, 1987; Lovitt & Jenkins, 1979; Ysseldyke, Algozzine, & Epps, 1983-84; Ysseldyke, Algozzine, Shinn, & McGue, 1982) assert that these figures represent an overidentification of students classified as learning disabled.

The major concern, according to Reynolds (1984-85), is the construction of a workable definition of learning disabilities. While Public Law 94-142 (1975) provides a definition to use in determining placement in a learning disability program, numerous experts in the field consider the definition superficial. While the definition provided in Public Law 94-142 (1975) gives some guidance in identifying students for learning disability services, the vagueness of the law is reflected by the development of a variety of definitions which are in current use (Adelman & Taylor, 1986). However, several experts (e.g., Algozzine, Forgnone, Mercer, & Triffiletti, 1979; Berk, 1984; O'Shea & Valcante, 1986; Reynolds 1984-85; Zimmerman & Williams, 1982) agree that the one common indicator in all the definitions is the existence of a severe discrepancy between aptitude and achievement. Even though the importance of the establishment of a severe discrepancy was immediately obvious, just as obvious was the subjective nature of developing a model to identify a severe discrepancy.

The criterion for identifying which individuals will be classified as learning disabled is determined mainly by the particular model of severe discrepancy employed. According to Reynolds (1984-85), and Epps, Ysseldyke, and Algozzine (1983) the differences in identification can vary in a random sample of the

population from 1% to more than 30% depending on the model being used. The "lack of a specific definition, improper or lack of application of the severe discrepancy criterion, and the failure to develop appropriate mathematical models (with reference to the severe discrepancy criterion) are the primary, certainly interrelated, causes of these disparities" (Reynolds, 1984-85, p. 455) in identification.

Another variable that contributes to the inconsistent diagnosis of a severe discrepancy between aptitude and achievement is the aptitude level of the student. Numerous researchers (e.g., Reynolds, 1984-85; Shepard, 1980; Willson & Reynolds, 1984-85; Wilson, Cone, Bush & Allee, 1983) agree that students with high aptitude scores will be overidentified and students with low aptitude scores will be underidentified unless regression to the mean is controlled. Wilson and Cone (1984) explained regression as "the tendency of extreme scores on one measure to be less extreme on a second related measure. The second scores regress or 'move' toward their mean" (p. 97). Shepard (1980) stated that while the regression phenomenon is recognized by statisticians and measurement theorists it appears to be ignored in much of the literature concerning the quantification of a severe discrepancy for learning disabilities identification. Therefore the aptitude level of the student can contribute to the probability of the identification of a severe discrepancy.

In addition to model and aptitude level, the grade level of the student can contribute to the inconsistency of identification of a severe aptitude-achievement discrepancy (Hanna, Dyck, & Holen, 1979; Shepard, 1980; Wilson et al., 1983). A student's achievement does not follow a straight line growth curve but tends to flatten at the upper grade levels. Reynolds (1981) noted that there is little change on standard scores on achievement tests between students who are two years below grade placement and students who are on grade level. The rate of learning is not constant throughout the years and academic concepts are acquired at different rates. Since, some subjects, such as reading, are not taught at the upper grade levels, these scores can only represent estimations from earlier grades. Shepard (1980) purports that the use of standard score models would correct for the grade level effect.

Since the various models, the aptitude level of the student, and the grade level of the student tend to produce conflicting results in the identification of a severe discrepancy, a careful examination of these

variables is necessary. Forness, Sinclair, and Guthrie (1983) reported that a larger percentage of students are currently being identified as learning disabled than any other exceptionality in special education; therefore, the practices and policies of identification are being examined. School personnel spend an excessive amount of time attempting to define, identify, classify, and recommend children for placement in learning disabled programs. Frequently, little professional agreement is achieved despite all this examination and study. This disagreement among professionals has created a disparity in the percentage of students being identified as learning disabled (Algozzine and Ysseldyke, 1986).

Even though there are a number of different procedures for identifying students as learning disabled, all of the procedures seem to have a common objective. Various authors (e.g., Algozzine, Ysseldyke, & Shinn, 1982; Chalfant & King, 1976; Forness et al., 1983; O'Donnell, 1980) indicate that the identification of a severe discrepancy between an individual's aptitude and achievement scores is the common objective for diagnosis of leaning disabilities. "The discrepancy concept is central to a learning disability diagnosis since it serves to distinguish learning disabilities from other conditions such as mental retardation or underachievement" (Forness et al., 1983, p. 107).

The first step in the development of an effective definition of learning disabilities is to substantiate that a procedure exists which quantifies a severe discrepancy. This step can be accomplished through a systematic analysis of the current models. The models can be analyzed and the effectiveness of the various procedures documented. The frequency of occurrence in the discrepancy scores will need to be considered also. A discrepancy should occur relatively infrequently in the normal population to avoid overidentification. Establishing documented effective aptitude-achievement discrepancy models will be quite useful in helping to identify individuals with learning disabilities.

However, as Reynolds (1984-85) stated, establishing the existence or a severe discrepancy does not automatically diagnose a student as learning disabled. The diagnosis and recommendation of placement in a learning disabilities program can only be made by an appropriate multidisciplinary team as outlined in the procedures of the rules and regulations implementing Public Law 94-142 (1975). Even though the establishment of a severe discrepancy is a minimal requirement, according to the law it is a necessary

requirement before an individual can be considered for a diagnosis of learning disability. Therefore, there is a need for rational guidelines for the identification of a severe discrepancy since it constitutes a major component in the determination of placement in a learning disability program.

As a result of this need, numerous formulas for calculating a severe discrepancy have been proposed. The formulas that have been proposed can be classified into two categories. These categories are nonstandard score models and standard score models. Few studies have evaluated the comparability of the models and limited evidence is available to substantiate the validation of any model. An empirically validated model will be the first step toward enhancing the sophistication in the field of learning disabilities (Cone & Wilson, 1981).

According to Wilson et al. (1983), nonstandard score models are "the oldest but least defensible methods of quantifying academic discrepancy" (p. 241). However, as Reynolds (1985) commented, these models remain immensely popular despite their serious psychometric deficiencies. A review reveals that nonstandard score models typically consists of a combination of two or more of the following components: chronological age, intelligence quotient, mental age, years in school, actual achievement, grade age, and grade placement. The achievement score in the discrepancy formulas is typically identified as a grade equivalent score on an achievement test and the aptitude score is the mental age or the intelligence quotient. It is the similarity of treating these components as interval and ratio scale measurements that constitute the limitations of nonstandard score models. As numerous experts (e.g., Cone & Wilson, 1981; Hanna et al., 1979; Reynolds, 1985; Wilson et al., 1983) have stated, age and grade equivalents cannot legitimately be subjected to the subtraction, addition, division, and multiplication computations, used by these methods because of the lack of a consistent unit of measurement. Therefore, the use of these formulas can be rejected as inadequate and misleading. Only standard score models that possess equal intervals of measurement can be considered as possible methods of quantifying a severe discrepancy.

There are three standard score discrepancy formulas that merit attention. These formulas are Model One: Simple Difference; Model Two: Estimated True Difference; and Model Three: Frequency of Regression Prediction. The three models are outlined in Table 1.

Table 1. Three Models for Determining a Severe Discrepancy

Model One: Simple Difference	
Calculate:	$Z_x = (X - \bar{X}) / S_x$ and $Z_y = (Y - \bar{Y}) / S_y$
Subtract:	$D = Z_x - Z_y$
Divided by Standard Error:	$Z_{Diff} = D / [2 - r_{xx} - r_{yy}]^{1/2}$
Compare Z to:	$Z_a = Z_{.05} = 1.65$
Model Two: Estimated True Difference Score	
Calculate:	$Z_{t_x} = r_{xx} Z_x$ and $Z_{t_y} = r_{yy} Z_y$
Subtract:	$D = Z_{t_x} - Z_{t_y}$
Divided by Standard Error:	$Z_{Diff} = D / [r_{xx}^2 (1 - r_{xx}) + r_{yy}^2 (1 - r_{yy})]^{1/2}$
Compare Z to:	$Z_a = Z_{.05} = 1.65$
Model Three: Frequency of Regression Prediction	
Calculate:	$Z_{\hat{y}} = r_{xy}(Z_x)$
Subtract:	$D = Z_{\hat{y}} - Z_y$
Divided by Standard Error:	$Z_{Diff} = D / [1 - r_{xy}^2]^{1/2}$
Compare Z to:	$Z_a = Z_{.05} = 1.65$

Note: X is an individual's aptitude score; $\bar{X}$ is the mean aptitude score; S_X is the standard deviation of the aptitude score; Y is the individual's achievement score; $\bar{Y}$ is the mean achievement score; S_Y is the standard deviation of the achievement score; r_{XX} is the internal consistency reliability coefficient for the aptitude test; r_{YY} is the internal consistency reliability coefficient for the achievement test; Z_{Tx} is the estimated true aptitude standard score; Z_{Ty} is the estimated true achievement standard score; $Z_{\hat{y}}$ is the predicted achievement standard score; r_{XY} is the correlation coefficient between the aptitude and achievement scores.

Model One: Simple Difference presents an objective means of documenting a discrepancy between aptitude and achievement compared to the nonstandard score models. In this context, Model One: Simple Difference is of considerable value. Forness et al. (1983) compared the model to seven nonstandard score models. The study yielded striking disparate results in terms of the number of children identified as having a severe discrepancy by each formula. Three other studies conducted by Cone and Wilson (1981), Dominick (1979) and Erickson (1975) also found that Model One: Simple Difference and nonstandard score models identified different students as having a severe discrepancy. Additionally, though, these three studies indicated that the students identified by Model One had significantly higher IQ scores than the students identified by the nonstandard score models.

While Model Two: Estimated True Difference Score has been studied extensively in test theory, little research has compared the model with other models in identifying a severe discrepancy. One study, though, by Bennett and Clarizio (1988) compared Model Two to a simple difference model and found that the models select similar proportions of students as having a severe discrepancy when using the Peabody Individual Achievement Test (PIAT) and the Kaufman Assessment Battery for Children (K-ABC). In fact, the authors of the study concluded that the simple difference model and the estimated true difference model could be used interchangeably in the calculation of a severe discrepancy when tests have high reliabilities.

In spite of the limited research studies conducted using Model Two, Berk (1984) had advocated the true difference score method as a meaningful and statistically defensible definition of a severe discrepancy. However, Braden (1987), Reynolds (1984-85), and Shepard (1980), have documented the inadequacy of the model to account for regression effects in the relationship between aptitude and achievement. They believe that regression should be a primary consideration in the development of an aptitude-achievement discrepancy model and that the use of Model One and Model Two will result in a bias toward individuals diagnosed as having a severe discrepancy.

The simple difference and the estimated true difference models are based on the assumption that an individual's z score on an aptitude measure will be the same as the individual's z score on an

achievement measure. This assumption would be accurate if the correlation between aptitude and achievement were perfect ($r = 1.0$ or $r = -1.0$). Since the actual correlation is not perfect (usually $r =$ about .6), there will be a regression to the mean. "This is not theory; it is an unavoidable factual consequence of the positive correlation between aptitude and achievement" (Reynolds, 1984-85, p. 461). Therefore, any model to determine the existence of a severe discrepancy that does not account for regression effects will identify more individuals with above average ability than individuals with below average ability.

According to Reynolds (1984-85) Model Three: Frequency of Regression Prediction properly addresses the regression between aptitude and achievement. The model uses the differences between the predicted achievement and the actual achievement while both Model One and Model Two ignore the predicted achievement scores. Therefore, Model Three should identify fewer discrepancy scores in the high IQ levels than Models One and Two identify. Research conducted by Braden (1987) supports the regression theory. He used two formulas to calculate discrepancies, the simple difference method and a regression method. The results indicated that the probability of an identification of a severe discrepancy varied drastically across intelligence intervals for the simple difference method, but the probability remained constant across intelligence levels for the regression method. The simple difference method identified more discrepant scores in the higher IQ level than in the lower IQ level.

However, Valus (1986) reported different results in her study. A severe discrepancy was computed for 88 subjects using a simple difference model and 72 subjects using a regression model. The comparison found no significant differences in the two procedures for identifying a severe discrepancy. One factor contributing to the conflicting results is the lack of a statistical criterion for determining a cut off score for either model. The differences were compared to a regressed standard score table. The regression table values adjusted for the differences of the simple difference model to account for the regression effect. Also, low IQ scores were over represented in the study. According to Reynolds (1984-85), regression models identify fewer students with high aptitudes as having a severe discrepancy than other standard score

models identify. Since this study did not include students with high aptitude scores, differences between the two models were minimized.

Bennett and Clarizio (1988) also compared regression models in Michigan with other standard score models. Four formulas were used to compare 86 students's test scores on the K-ABC, WISC-R, and PIAT. The four methods were a simple difference model, an estimated true score difference, a regression model, and a regression model that included an adjustment for unreliability. The regression model identified significantly fewer and different proportions of students than the simple difference model, the estimated true difference model, and the adjusted regression model identified. The number of discrepant scores identified by the regression model adjusted for unreliability was dependent on the reliabilities of the tests. This dependency indicated that test with low reliabilities selected appreciably less discrepant scores than did tests with high reliabilities. Also, scores from different tests influenced the identification process. Scores from the PIAT identified significantly fewer students than scores from the K-ABC without regard to the formula used to calculate the severe discrepancy.

While standard score models have been studied extensively in test theory, research evidence on the models is sparse. The available research has indicated that the specific test administered, students' IQ levels, students' grade levels, and formulas utilized have influenced the probability of identifying a severe discrepancy. One theme that has permeated the evaluation of aptitude-achievement discrepancy models is that important decisions about students must be based on reliable and valid information.

Purpose of the Study

This study compared the results of applying three aptitude-achievement discrepancy models to a sample of students' aptitude and achievement scores in grades 2, 3, 4, and 5 who had been referred for psychoeducational testing in several school systems in the Southeast. Specifically, the purpose of the study was to determine whether different groups of students (i.e., groups classified by grade level and aptitude level) were identified as having a severe discrepancy according to the formula applied.

Method

The 580 subjects for the study were randomly selected from a population of 1,452 students, grades two through five, who were referred for psychoeducational testing from 1986 to 1990 and had WISC-R Full Scale IQ scores and K-TEA Comprehension Form Battery Composite scores available. The subjects were selected from seven urban and suburban school systems in the Southeast. There were 150 students selected in each grade level two through four and 50 students in each aptitude level (low, average, high) at each grade level. Grade five had 130 student and 50 students in low and average aptitude levels and 30 students in the high aptitude level. For the purposes of this study high aptitude included test scores of 111 or above, average aptitude included test scores of 90 to 110, and low aptitude included test scores of 70 to 89 as measured by the WISC-R Full Scale IQ score. As indicated in Table 2, students with low aptitudes were older than students with average and high aptitudes across all grade levels. Students with average aptitudes were older than students with high aptitudes across all grade levels.

Table 2. Ages of Subjects When Tested

Grade	Aptitude Level	Mean Age	Standard Deviation
2	Low	8.50	.81
2	Average	7.98	.57
2	High	7.77	.47
3	Low	9.68	.81
3	Average	9.26	.57
3	High	8.90	.55
4	Low	10.51	.70
4	Average	10.19	.89
4	High	10.06	.44
5	Low	11.72	1.10
5	Average	11.17	.94
5	High	10.90	.79

Instrumentation

Two instruments were used in this study; the Wechsler Intelligence Scale for Children--Revised (WISC-R) and the Kaufman Test of Educational Achievement, Comprehensive Form (K-TEA). The WISC-R is a measure of intelligence (Wechsler, 1974) and the K-TEA is an individual test of academic achievement (Kaufman & Kaufman, 1985). These two instruments were chosen to be used in the study because of their application for use in identifying learning disabled students.

Wechsler Intelligence Scale for Children--Revised. The WISC-R is described in the manual "as a test of general intelligence" (Wechsler, 1974, p. iii). It is an individually administered test given by a trained psychometrist to children ages 6 years 0 months through 16 years 11 months. Testing time is approximately 50 to 75 minutes. The instrument consists of three scales; the Verbal Scale, the Performance Scale, the Full Scale IQ. The Verbal Scale is composed of six subtests, but only five are mandatory. The Performance Scale is arranged in a similar manner with five mandatory tests and one supplemental test. The Full Scale Verbal and Performance yield an IQ with a mean of 100 an a standard deviation of 15 (Wechsler, 1974).

According to the WISC-R Manual, the standardization sample included a total of 2,200 cases with 100 boys and 100 girls in eleven age groups ranging from 6 1/2 through 16 1/2 years. The 1970 U.S. census was used in the stratification of the sample in regard to race, geographic region, occupation of head of household, and urban-rural residence (Wechsler, 1974).

The split-half reliability coefficients for each subtest were obtained with the appropriate correction for the full length of the test by the Spearman-Brown formula (Wechsler, 1974). A measure of internal consistency was provided by the split-half procedure. Test-retest or stability coefficients were computed for these two subtests (Wechsler, 1974). As reported in the WISC-R Manual, the average reliability coefficients are .94, .90, and .96 for the Verbal, Performance, and Full Scale IQs, respectively. The individual subtest reliabilities are satisfactory with the average coefficient for the Performance Scale, ranging form .70 to .85 and the Verbal Scale, ranging from .77 to .86 (Wechsler, 1974).

Kaufman Test of Educational Achievement, Comprehensive Form. The K-TEA is an "individually administered measure of the school achievement of children and adolescents in grades 1 through 12" (Kaufman & Kaufman, 1985, p. 1). The test is administered individually by teachers, psychologists, counselors, or others who are trained in educational or psychoeducational testing. Test time varies because none of the items on the test are timed. However, the approximate testing time ranges from 20 minutes to 65 minutes (Kaufman & Kaufman, 1985).

There are five subtests of the K-TEA Comprehensive Form including Mathematics Application, Mathematics Computation, Reading Decoding, Spelling, and Reading Comprehension. A Mathematics Composite score is computed from the two subtests; Mathematics Application and Mathematics Computation. The Reading Composite is derived from scores on the subtests Reading Decoding and Reading Comprehension. All five subtest scores are included in the Battery Composite. Two sets of standard scores are available, age-based and grade-based scores. Standard scores with a mean of 100 and a standard deviation of 15 are reported for all subtests and Composite scores (Kaufman & Kaufman, 1985).

Two standardization samples were documented, one during the spring of 1983, and one during the fall of 1983, to accurately measure a student's educational performance at the beginning and ending of the school year. The standardization program in the spring and fall included more than 1,400 and 1,000 students respectively. The proportional composition of the sample closely matched the 1980 U.S. population on the following stratification variables: grade, sex, geographic region, socioeconomic level (as reflected by parent level of education), and race or ethnic group. Split-half reliability coefficients by age group and grade level for the five subtests and battery composites ranged from .83 to .98. The test-retest reliability coefficient for the separate subtests ranged from .83 to .96 and the coefficients for the three composites ranged from .90 to .97 (Kaufman & Kaufman, 1985).

Evidence of construct validity was documented through internal analyses, developmental changes and internal consistency. The K-TEA demonstrated both age and grade differentiation to support developmental changes. The internal consistency was determined by correlating the subtest scores with the Battery Composite. The coefficient for the subtest and composite scores by grade level ranged from .65 to

.91. The concurrent validity was documented through external analyses. Correlation coefficients between the K-TEA and three group achievement tests ranged from .67 to .90. Content validity was indicated by presenting a rationale for the subtests. The rationale involved a description of attention given to item development, tryouts, and analysis. However, the manual warns that there is an insufficient floor for all subtests for students in grade one who are below average (Kaufman & Kaufman, 1985).

Procedures for Data Collection

School systems were selected based on availability and accessibility of test scores and to include a diverse population. Initial contact was made to school systems that routinely administered the WISC-R and the K-TEA to referred students. Data were collected from the systems that granted permission to conduct the study. After parental consent was granted, all test scores were recorded except for students whose aptitude estimates fell into the educable mentally impaired range (IQ > 70) because these scores were not included in the study. After the data were recorded, the subjects were randomly selected for inclusion in this study.

The WISC-R had been administered to the participants by certified school psychometrists or school psychologists. The K-TEA had been administered by special education teachers, school counselors, school psychometrists, or school psychologists. These tests were administered in differing order within an interval of 1 day to 6 weeks. There is assumed to be no practice effect between the two tests to influence the interpretation of the results. Other variables such as educational experiences and maturation were controlled by a reasonably short time interval between tests.

Hypotheses, Data Analysis, and Results

A factorial two between- and one within-subjects with repeated measures analysis of variance design (ANOVA) was used for this study, a design that Kennedy (1978) and Kirk (1982) reported to be appropriate for investigations involving between-and within-subjects variables. The independent variables A and B, grade level and aptitude level respectively, crossed with the independent variable C, discrepancy model. The assigned value of one indicating the existence of a severe discrepancy or zero indicating that

the discrepancy was not severe as determined by each of the three models was the dependent variable. Also, subjects (S) were regarded as a variable since the design incorporates a within-subjects factor.

A random sample of subjects from seven school systems was assigned among the combinations levels of the variables A and B, with the restriction that there be 50 sets of test scores in each AB combination, except grade five high aptitude which had 30 sets of scores. Each set of test scores (WISC-R Full Scale IQ score and K-TEA Composite score) was tested by each formula (variable C) to determine if there was a severe discrepancy between aptitude and achievement at the .05 level of significance ($p \leq .05$). The mean internal consistency reliability coefficient (r_{XX}) for the WISC-R Full Scale IQ of .96 as reported by the WISC-R manual was used in the calculation of each of the three models (Wechsler, 1974). A mean by grade level of the split-half reliability coefficient of .98 (r_{YY}) was used for the K-TEA internal consistency reliability as reported by Kaufman and Kaufman (1985). A Pearson correlation coefficient of .663 (r_{XY}) between the WISC-R Full Scale IQ score and the K-TEA Composite score for the 580 sets was calculated and used in the formula for Model Three: Frequency of Regression Prediction.

The discrepancies for each model were computed. The subjects identified as having a severe discrepancy were assigned a score of one and subjects not identified as having a severe discrepancy were assigned a score of zero. Therefore, the recorded data were dichotomous. The use of dichotomous data in analysis of variance (ANOVA) procedures is appropriate according to Bruning and Kintz (1987). The actual computational steps for dichotomous data are the same as for discrete data the only difference is that the data recorded for each subject were one or zero.

The means reported for each model represent the percentage of students identified as having a severe discrepancy between aptitude and achievement. The means and standard deviations for each model by grade level and aptitude level are listed in Table 3. Model Three: Frequency of Regression Prediction had smaller means than Model One: Simple Difference and Model Two: Estimated True Difference at all grade levels and aptitude levels. The largest differences among the means were evident at Grade 5 High Aptitude. Model One and Two identified 90% of the students as having a severe discrepancy and Model Three only identified 20% of the students.

Table 3. Sample Size, Mean, and Standard Deviation of the Severe Discrepancy Scores

Grade	Aptitude Level	n	Model 1 M	Model 1 SD	Model 2 M	Model 2 SD	Model 3 M	Model 3 SD
2	Low	50	.18	.39	.26	.44	.02	.14
	Average	50	.72	.45	.72	.45	.16	.37
	High	50	.84	.37	.84	.37	.20	.40
3	Low	50	.22	.41	.30	.46	.06	.24
	Average	50	.66	.48	.68	.47	.14	.35
	High	50	.72	.45	.72	.45	.20	.40
4	Low	50	.24	.43	.28	.45	.02	.14
	Average	50	.50	.51	.54	.50	.14	.35
	High	50	.74	.44	.74	.44	.16	.37
5	Low	50	.38	.49	.38	.49	.10	.30
	Average	50	.54	.50	.56	.50	.14	.35
	High	30	.90	.31	.90	.31	.20	.41

Note: Low Aptitude = WISC-R Full Scale IQ scores 70 - 89; Average Aptitude = WISC-R Full Scale IQ scores 90 - 110; High Aptitude = WISC-R Full Scale IQ scores 111 and above.

The seven null hypotheses used to guide this investigation were tested hierarchically. A significant interaction of Hypothesis 1 required analysis of Postulate 1 and nonsignificant interaction required analysis of Hypotheses 2, 3, and 4. A significant interaction of any of these three hypotheses required testing of the hypotheses' paired postulate and a nonsignificant interaction of the three hypotheses required testing of Hypotheses 5, 6, and 7. Any significant differences required further analysis of the hypothesis' paired postulate.

ANOVA procedures were used to test significant Fs for each of the seven postulates when required according to the hierarchical design. Tukey's HSD (honestly significant difference) test with harmonic mean was the follow-up analysis for each of the postulates that required testing.

Hypothesis 1

There is no significant interaction between grade level, aptitude level, and discrepancy model for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 1: If a statistically significant interaction exists, then there will be no statistically significant differences among cell means.

Hypothesis 2

There is no significant interaction between grade level and discrepancy model for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 2: If a significant two-way interaction exists between grade level and discrepancy model, then there will be no statistically significant differences among cell means.

Hypothesis 3

There is no significant interaction between aptitude level and discrepancy model for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 3: If a statistically significant two-way interaction exists between aptitude level and discrepancy model, then there will be no statistically significant differences among cell means.

Hypothesis 4

There will be no statistically significant two-way interaction between aptitude level and grade level for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 4: If a statistically significant two-way interaction exists between aptitude level and grade level, then there will be no statistically significant differences among cell means.

Hypothesis 5

There will be no statistically significant main effect difference among grade levels for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 5: If statistically significant main effect differences exists among grade levels, then there will be no statistically significant differences among main effect cell means.

Hypothesis 6

There will be no statistically significant main effect difference among aptitude levels for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 6: If statistically significant main effect differences exists among aptitude levels, then there will be no statistically significant differences among main effect cell means.

Hypothesis 7

There will be no statistically significant main effect difference among discrepancy models for students classified or not classified as having a severe discrepancy between aptitude and achievement.

Postulate 7: If statistically significant main effect differences among discrepancy models exist, then there will be no statistically significant differences among main effect cell means.

The three-way interaction among grade level, aptitude level, and discrepancy model and the two-way interactions between grade level and discrepancy model and between aptitude level and grade level were nonsignificant. Therefore, null hypotheses 1, 2, and 4 were not rejected and their paired postulates were not tested. Hypotheses 5, 6, and 7 were not tested since Hypothesis 3 was significant, $p \leq .05$.

The two-way interaction between aptitude level at discrepancy model was significant at the .05 level. Therefore, null hypothesis 3 was rejected and its paired postulate tested. An ANOVA of aptitude and model resulted in significant F values, $p \leq .05$, for Aptitude at Model One, Model Two, Model Three and Model at Low Aptitude, Average Aptitude, High Aptitude. Tukey's HSD with harmonic mean procedure was used to examine Aptitude at Model and Tukey's HSD procedure was used to examine Model at Aptitude.

Three statistically significant difference values were obtained for Model One at High, Average, and Low Aptitudes. Use of Model Two also resulted in three significant difference values at High, Average, and Low aptitudes. In contrast, differences found using Model Three resulted in only one statistically significant difference value between High aptitude and Low aptitude.

In addition to the various statistically significant F values, $p \leq .05$, practical significance was investigated using eta squared. The computation of the formula indicated that 30% of the variance observed in the study can be associated with the manipulation of the three models and aptitude levels of the students. According to Kennedy (1978), "significant efforts reported in the educational literature typically do not explain more than 10% of the variance" (p. 221); therefore, practical significance was found.

Conclusions

Based on the findings in this study general conclusions are stated below:

1. Model One: Simple Difference had significant differences across all aptitude levels suggesting that the model tends to select different proportions of students with low, average, and high aptitudes.

2. Model Two: Estimated True Difference had significant differences across all aptitude levels suggesting that the model tends to select different proportions of students with low, average, and high aptitudes.

3. Model Three: Frequency of Regression Prediction model identified significantly more discrepancies for students with high aptitudes then students with low aptitudes in contrast to the theory presented in the literature. The use of grade-based standard scores instead of age-based standard scores could explain the contradictory results.

4. Model Thee: Frequency of Regression Prediction identified similar proportions of discrepancies for students with high aptitudes and average aptitudes.

5. Models One and Two selected similar proportions of students across the three aptitude levels. These two models might be used interchangeably in the calculation of a discrepancy.

6. Model Three selected considerably fewer discrepancies across the three aptitude levels than the other two models. Therefore, Model Three cannot be used interchangeably with Models One and Two.

Implications and Discussion

The lack of agreement between a regression model and other standard score models in identifying students with a severe discrepancy between ability and achievement suggests that the model chosen could

have a significant impact on the identification process. If a regression model is chosen as a discrepancy criterion, fewer students will be identified than would be identified using a simple difference model or an estimated true difference model. A regression model is not interchangeable with a simple difference or estimated true difference model; however, the simple difference and the estimated true difference models can be used interchangeable.

A study conducted by Bennett and Clarizio (1988) found comparable results. A regression model identified significantly fewer and different proportions of students than a simple difference model and estimated true difference model. Braden (1987) also found that a simple difference model identified significantly different students as having discrepancies than a regression model identified.

While Model Three: Frequency of Regression Prediction may control for the regression effect the use of grade-based standard scores could reduce the model's effectiveness. Theoretically, a regression model should not identify more aptitude-achievement discrepancies for high aptitude levels than low aptitude levels because the model purports to correct for the regression to the mean effect (Berk, 1984; Glass & Hopkins, 1984; Shepard, 1980; Reynolds, 1984-85; Wilson & Cone, 1984). The results from this study do not support this claim. One explanation for the contradictory results of this study was the use of grade-based standard scores instead of age-based standard scores for the K-TEA Composite Score. According to the descriptive statistics reported in Table 2, students with low aptitudes were older than students with average or high aptitudes across all grade levels. Also, students with average aptitudes were older than students with high aptitudes. The use of grade-based standard scores instead of age-based standard scores would result in similar K-TEA Composite Scores for students with low and average aptitudes and lower K-TEA Composite Scores for students with high aptitudes. Consequently, students with high aptitudes would have a greater chance than students with low and average aptitudes of having a discrepancy between aptitude and achievement using grade-based standard scores instead of age-based standard scores. Therefore, caution should be exercised with this model when identifying discrepancies for students using grade-based standard scores. Students with high aptitudes who are younger than their grade peers may be overidentified.

Models One and Two identified significantly different proportions of students across all aptitude levels; however, Model Thee maintained a consistent identification process for high and average aptitudes. Since Model Three: Frequency of Regression was more effective than Models One and Two in reducing the number of students identified as having a severe discrepancy and maintaining a consistent identification process across aptitude levels, this model is recommended.

For ease in computation of the severe discrepancies, practitioners should use a computer software package. There is a software package available for Model Three: Frequency of Regression Prediction (Reynolds & Stowe, 1985).

Based on the results of this study, the recommendations listed below merit further studies.

1. Similar analysis with other regression models should be conducted. Bennett and Clarizio (1988) compared two regression formulas, Model Three: Frequency of Regression Prediction and an adjusted regressed difference model. These two models selected different proportions of students.

2. An analysis to compare grade-based standard scores and age-based standard scores would be enlightening. One explanation given for the failure of Model Three: Frequency of Regression Prediction to control for the regression effect was the use of grade-based standard scores. If age-based standard scores had been used, the students with high aptitudes could have received higher achievement scores. These higher achievement scores would have resulted in a smaller difference or no difference between aptitude and achievement. Since the study resulted in limited support for the theoretical bases of Model Three: Frequency of Regression Prediction and the research reported by Braden (1987), further investigations are warranted.

3. A comparison of age-based achievement scores and age-based aptitude scores would provide valuable information. Hanna et al. (1979) advocates the need to compare aptitude and achievement test scores with the same reference groups. Different reference groups produce minor changes when students' ages are average for their grade level; however, the changes are severe when the students' ages are markedly different from their grade peers. The achievement test scores used in this study were grade-based scores and the aptitude test scores were age-based scores; therefore, the reference groups were

dissimilar. Also, low-aptitude students were on the average considerably older than high-aptitude students. Since the study compared different reference groups with extreme differences in ages for each grade level, additional comparisons are justified.

4. It would be valuable to replicate this study using different tests. Reliability coefficients of the tests are used in the discrepancy models; therefore, generalizability of the results of the study cannot be extended to other tests. Since school systems use different tests for the identification process, further study would be valuable.

References

Adelman, H. S., & Taylor, L. (1986). The problems of definition and differentiation and the need for a classification schema. Journal of Learning Disabilities, 19, 514-520.

Algozzine, B., Forgnone, C., Mercer, C., & Triffiletti, J. (1979). Toward defining discrepancies for specific learning disabilities: An analysis and alternatives. Learning Disability Quarterly, 2(4), 25-31.

Algozzine, B., & Ysseldyke, J. (1983-84). Learning disabilities as a subset of school failure: The over-sophistication of a concept. Exceptional Children, 40, 242-246.

Algozzine, B., & Ysseldyke, J. (1986). The future of the LD Field: Screening and diagnosis. Journal of Learning Disabilities, 19, 394-398.

Algozzine, B., Ysseldyke, J., & Shinn, M. (1982). Identifying children with learning disabilities: When is a discrepancy severe? Journal of Experimental Education, 40(4), 1-5.

Bennett, D. E., & Clarizio, H. F. (1988). A comparison of methods for calculating a severe discrepancy. Journal of School Psychology, 26, 359-369.

Berk, R. A. (1984). Screening and diagnosis of children with learning disabilities. Springfield, IL: Charles C. Thomas.

Braden, J. P. (1987) A comparison of regression and standard score discrepancy methods for learning disabilities identification: Effects on racial representation. Journal of School Psychology, 25, 23-29.

Bruning, J. L., & Kintz, B. L. (1987). Computational Handbook of Statistics (3rd ed.). Glenview, IL: Scott, Foresman and Company.

Chalfant, J. C., & King, F. S. (1976). An approach to operationalizing the definition of learning disabilities. Journal of Learning Disabilities, 9, 228-243.

Clarizio, H. F., & Phillips, S. E. (1986a). Sex bias in the diagnosis of learning disabled students. Psychology in the Schools, 23, 44-52.

Clarizio, H. F., & Phillips, S. E. (1986b). The use of standard scores in diagnosing learning disabilities: A critique. Psychology in the Schools, 23, 380-387.

Cone, T. E., & Wilson, L. R. (1981). Quantifying a severe discrepancy: A critical analysis. Learning Disability Quarterly, 4, 359-371.

Dominick, N. A. (1979). A comparison of techniques for the identification of children with reading disabilities. Unpublished doctoral dissertation, The University of North Carolina at Chapel Hill.

Epps, S., Ysseldyke, J., & Algozzine, B. (1983). Impact of different definitions of learning disabilities on the number of students identified. Journal of Psychoeducational Assessment, 1, 341-352.

Erickson, M. T. (1975). The z-score discrepancy method for identifying reading disabled children. Journal of Learning Disabilities, 8, 308-312.

Forness, S. R., Sinclair, E., & Guthrie, D. (1983). Learning disability discrepancy formulas: Their use in actual practice. Learning Disability Quarterly, 6, 107-114.

Frankenberger, W., & Harper, J. (1987). States' criteria and procedures for identifying learning disabled children: A comparison of 1981/82 and 1985/86 guidelines. Journal of Learning Disabilities, 20, 118-121.

Glass, G. V., & Hopkins, K. D. (1984). Statistical Methods in Education and Psychology (2nd ed.). New Jersey: Prentice-Hall.

Hanna, G. S., Dyck, N. J., & Holen, M. C. (1979). Objective analysis of achievement-aptitude discrepancies in LD classification. Learning Disability Quarterly, 2(4), 32-38.

Kaufman, A. S., & Kaufman, N. L. (1985). Comprehensive Form Manual for the Kaufman Test of Educational Achievement. Circle Pines, MN: American Guidance Service.

Kennedy, J. J. (1978). An introduction to the design and analysis of experiments in education and psychology. Lanham, MD: University Press of America.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, CA: Wadsworth, Inc.

Lovitt, T. C., & Jenkins, J. R. (1979). Learning disabilities research: Defining populations. Learning Disability Quarterly, 2, 46-50.

O'Donnell, L. E. (1980). Intra-individual discrepancy in diagnosing specific learning disabilities. Learning Disability Quarterly, 3, 10-18.

O'Shea, L. J., & Valcante, G. (1986). A comparison over time of relative discrepancy scores of low achievers. Exceptional Children, 53, 253-259.

Reynolds, C. R. (1981). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. Journal of School Psychology, 19, 350-358.

Reynolds, C. R. (1984-85). Critical measurement issues in learning disabilities. Journal of Special Education, 18, 451-476.

Reynolds, C. R. (1985). Measuring the aptitude-achievement discrepancy in learning disability diagnosis. Remedial and Special Education, 6(5), 37-48.

Reynolds, C. R., & Stowe, M. L. (1985). Severe discrepancy analysis operations manual. Bensalem, PA: TRAIN, Inc.

Shepard, L. (1980). An evaluation of the regression discrepancy method for identifying children with learning disabilities. Journal of Special Education, 14, 79-91.

United States Office of Education. (1977). Rules and regulations implementing Public Law 94-142 (The Education for All Handicapped Children Act of 1975.) Federal Register, 42.

Valus, A. (1986). Achievement-potential discrepancy status of students in LD programs. Learning Disabilities Quarterly, 9, 200-205.

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children (Revised) (WISC-R). NY: Psychological Corporation.

Willson, V. L., & Reynolds, C. R. (1984-85). Another look at evaluating aptitude-achievement discrepancies in the diagnosis of learning disabilities. Journal of Special Education, 18, 477-487.

Wilson, L. R., & Cone, T. (1984). The regression equation method of determining academic discrepancy. Journal of School Psychology, 22, 95-110.

Wilson, L. R., Cone, T. E., Bush, R., & Allee, T. (1983). A critique of the expectancy formula approach: Beating a dead horse? Psychology in the Schools, 20, 241-249.

Ysseldyke, J., Algozzine, B., & Epps, S. (1983-84). A logical and empirical analysis of current practice in classifying students as handicapped. Exceptional Children, 40, 160-166.

Ysseldyke, J. E., Algozzine, B., Shinn, M. R., & McGue, M. (1982). Similarities and differences between low achievers and students classified learning disabled. Journal of Special Education, 16, 73-85.

Zimmerman, D. W., & Williams, R. H. (1982). The relative error magnitude in three measures of change. Psychometrika, 47, 141-147.